Enhancing the BOADICEA cancer risk prediction model to incorporate new data on RAD51C, RAD51D, BARD1 updates to tumour pathology and cancer incidence

Andrew Lee,1,2 Nasim Mavaddat,1 Alex Cunningham,1 Tim Carver,1 Lorenzo Ficorella,1 Stephanie Archer,2 Fiona M Walter,2,3 Marc Tischkowitz,4 Jonathan Roberts,4 Juliet Usher-Smith,2 Jacques Simard,5 Marijanka K Schmidt,6 Peter Devilee,7 Vesna Zadnik,8 Hannes Jürgens,9 Emmanuelle Mouret-Fourme,10 Antoine De Pauw,10 Matti Rookus,11 Thea M Mooij,11 Paul PD Pharoah,1,2 Douglas F Easton,1,2 Antonis C Antoniou1

ABSTRACT

Background BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) for breast cancer and the epithelial tubo-ovarian cancer (EOC) models included in the CanRisk tool (www.canrisk.org) provide future cancer risks based on pathogenic variants in cancer-susceptibility genes, polygenic risk scores, breast density, questionnaire-based risk factors and family history. Here, we extend the models to include the effects of pathogenic variants in recently established breast cancer and EOC susceptibility genes, up-to-date age-specific pathology distributions and continuous risk factors.

Methods BOADICEA was extended to further incorporate the associations of pathogenic variants in BARD1, RAD51C and RAD51D with breast cancer risk. The EOC model was extended to include the association of PALB2 pathogenic variants with EOC risk. Age-specific distributions of oestrogen-receptor-negative and triple-negative breast cancer status for pathogenic variant carriers in these genes and CHEK2 and ATM were also incorporated. A novel method to include continuous risk factors was developed, exemplified by including adult height as continuous.

Results BARD1, RAD51C and RAD51D explain 0.31% of the breast cancer polygenic variance. When incorporated into the multifactorial model, 34%–44% of these carriers would be reclassified to the near-population and 15%–22% to the high-risk categories based on the UK National Institute for Health and Care Excellence guidelines. Under the EOC multifactorial model, 62%, 35% and 3% of PALB2 carriers have lifetime EOC risks of <5%, 5%–10% and >10%, respectively. Including height as continuous, increased the breast cancer relative risk variance from 0.002 to 0.010.

Conclusions These extensions will allow for better personalised risks for BARD1, RAD51C, RAD51D and PALB2 pathogenic variant carriers and more informed choices on screening, prevention, risk factor modification or other risk-reducing options.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Pathogenic variants in BARD1, RAD51C and RAD51D have recently been established as breast cancer susceptibility genes, and pathogenic variants in PALB2 have been shown to be associated with epithelial ovarian cancer risk. No cancer risk prediction model currently exists which incorporates these associations.

WHAT THIS STUDY ADDS

The BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) multifactorial breast and ovarian cancer risk prediction model has been extended to incorporate these associations and has been implemented in the CanRisk tool (www.canrisk.org) for use by healthcare professionals.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE AND/OR POLICY

The enhanced risk prediction models will enable healthcare professionals to provide personalised breast and epithelial ovarian cancer risks to BARD1, RAD51C, RAD51D and PALB2 pathogenic variant carriers and will allow for more informed choices on cancer risk management options.

INTRODUCTION

Breast cancer (BC) and epithelial tubo-ovarian cancer (EOC) are two of the most common cancers in women.1,2 Through mammography or other methods, screening for BC can reduce mortality, and organised screening is available in most developed countries.3 For EOC, no effective screening exists, but the disease can be prevented by salpingo-oophorectomy. However, these preventative options are associated with adverse effects. Therefore, identifying those at increased risk may help to target screening and preventative options to those most likely to benefit.3 Both BC and EOC...
risks are multifactorial diseases, with family history of cancer (FH), genetic factors and lifestyle, hormonal and reproductive risk factors (RFs) all contributing to risk.\textsuperscript{5-7}

Previously we developed the BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) model for BC risk prediction and for the likelihood of carrying pathogenic variants (PVs) in BC susceptibility genes. BOADICEA v5 incorporates the effects of PVs in five BC susceptibility genes (\textit{BRCA1, BRCA2, PALB2, CHEK2} and \textit{ATM}), the effects of known common genetic variants summarised as a polygenic risk score (PRS, accounting for \textasciitilde20\% of the polygenic variance), and a polygenic component that accounts for any residual familial aggregation.\textsuperscript{8-9} We also developed a similar EOC model (Ovarian Cancer Model v1) that considers the effects of PVs in \textit{BRCA1}, \textit{BRCA2}, \textit{RAD51D}, \textit{RAD51C} and \textit{BRIP1} on EOC together with a PRS (accounting for \textasciitilde5\% of the polygenic variance) and a residual polygenic component.\textsuperscript{10-12} BOADICEA includes mammographic density and both models incorporate the effects of known lifestyle, hormonal, reproductive and anthropometric RFs. In addition, the models incorporate breast tumour heterogeneity by considering the distributions of tumour oestrogen receptor (ER) and triple-negative (TN) (ER, progesterone receptor and human epidermal growth factor receptor 2 negative) status for \textit{BRCA1} and \textit{BRCA2} PV carriers and the general population.\textsuperscript{12} Both models are freely available to healthcare professionals via the CanRisk webtool (www.canrisk.org) and are widely used by healthcare professionals.\textsuperscript{14}

Recently, large population-based and family-based targeted sequencing studies have established that PVs in \textit{RAD51C}, \textit{RAD51D} and \textit{BARD1} are associated with BC risk,\textsuperscript{15-16} and that PVs in \textit{PALB2} are associated with EOC risk.\textsuperscript{17-18} In addition, analysis of the tumour characteristics in the BRIDGES study has provided age-specific estimates of the distributions of tumour characteristics for PV carriers in all established susceptibility genes.\textsuperscript{19} A further limitation of the previous models is that all epidemiological RFs are treated as categorical. However, some RFs (eg, height, body mass index (BMI) mammographic density) are intrinsically continuous, and discretisation results in a loss of information, reducing their predictive ability. Here we extend both models to explicitly model the effects of PVs in the recently established BC and EOC susceptibility genes and incorporate up-to-date age-specific pathology distributions. We present a methodological framework for incorporating continuous RFs into the model, and we demonstrate this by including height as a continuous variable. Finally, we describe updates to the population reference cancer incidence rates used in the models by incorporating more up-to-date incidences, incidences for additional countries and refining the derivation of birth-cohort-specific incidences for inclusion in the models that address sparsity in the population incidence data.

**METHODS**

**Rare moderate-risk pathogenic variants**

Both BOADICEA and the EOC models, model cancer incidence as an explicit function of PVs in known high- and moderate-penetrance susceptibility genes (major genes) together with a polygenic component.\textsuperscript{6-12} By using an explicit genetic model, they can account for both genetic testing and detailed FH. BOADICEA includes the genes \textit{BRCA1, BRCA2, PALB2, CHEK2 and ATM}, with dominance in that order, along with a BC susceptibility polygenic component. The EOC model includes the genes \textit{BRCA1, BRCA2, RAD51D, RAD51C} and \textit{BRIP1}, with dominance in that order, along with an EOC susceptibility polygenic component. Details of the underlying model are included in the online supplemental material. The values of the parameters for the original models were determined by complex segregation analysis.\textsuperscript{9} However, this was not possible for the extended versions since no sufficiently large dataset containing all the model features was available. Instead, we adopted a synthetic approach,\textsuperscript{23} in which additional model parameters are taken from large-scale external studies.\textsuperscript{8,13-21}

Here, BOADICEA was extended to explicitly model the effects of PVs in \textit{BARD1, RAD51C} and \textit{RAD51D} (i.e. in total, eight BC susceptibility genes), while the EOC model was extended to include \textit{PALB2} (i.e. in total, six EOC susceptibility genes). In both models, the effects of PVs were included as major genes and are parameterised by their allele frequency in the general population and their age-specific relative risks (RRs). The BC RR for carriers of PV in \textit{BARD1} was taken from the BRIDGES study,\textsuperscript{15} while those for \textit{RAD51C} and \textit{RAD51D} were the meta-analysed values from Dorling et al\textsuperscript{24} and Yang et al.\textsuperscript{25} The EOC RR for \textit{PALB2} PV carriers was taken from Yang et al.\textsuperscript{15} The BRIDGES study\textsuperscript{15} suggested that the RR estimates associated with PVs in ATM are lower than the previously assumed estimate of 2.8,\textsuperscript{21} and it was therefore updated to the Dorling et al\textsuperscript{25} estimate. The previously assumed RR estimates for PVs in \textit{BRCA1, BRCA2, PALB2} and \textit{CHEK2}\textsuperscript{24,21} were based on large studies that enabled the estimation of age-specific risks or were estimated as part of the BOADICEA model fitting process, and were not updated, except for the \textit{BRCA2}-associated EOC RRs for ages 59 and over (online supplemental material). The PV carrier frequencies for \textit{PALB2, CHEK2} (including all PVs), ATM, \textit{BARD1, RAD51D, RAD51C} and \textit{BRIP1} and screening test sensitivities for all genes were derived from Dorling et al.\textsuperscript{15} We used the BRIDGES study to derive these frequency estimates as it is a very large population-based dataset that includes targeted sequencing data. Frequencies were based on the control frequencies in European populations, adjusted for the assumed sensitivity of the sequencing and the fact that large rearrangements were not detectable (online supplemental material). The default sensitivities were then calculated, assuming that clinical genetic testing will detect all known pathogenic mutations except for large rearrangements (except \textit{BRCA1} and \textit{BRCA2}, where testing for large rearrangements is routinely done). All model parameters for PVs are given in table 1.

As the polygenic component captures all residual familial aggregation not explained by the major genes, the previous models implicitly included the contributions of PVs in the new genes (ie, \textit{BARD1, RAD51C} and \textit{RAD51D} for BOADICEA and \textit{PALB2} for the EOC model). Therefore, to avoid double counting their contribution, it was necessary to remove their contribution from the polygenic component by adjusting the log-RR per SD of the polygenic component such that the total variance of the polygenic component and the new genes is the same as that of the polygenic component of the previous model\textsuperscript{25} (online supplemental material).

The association between \textit{PALB2} PVs and EOC was also included in the BOADICEA model, and the associations with male BC and pancreatic cancer have been included in both models.\textsuperscript{15}

The impact of including PVs in the new BC susceptibility genes on risk prediction were assessed by considering the risk categories described in the National Institute for Health and Care Excellence familial BC guidelines\textsuperscript{26} for hypothetical women with different ages or family history. For lifetime risk (aged 20–80 years), three categories are defined: (1) near-population risk, for
Table 1  Parameters used to include the effects of rare high-risk and intermediate-risk pathogenic variants in the models

<table>
<thead>
<tr>
<th>Gene</th>
<th>Allele Freq</th>
<th>SS</th>
<th>RR of female breast cancer (95% CI)</th>
<th>RR of EOC (95% CI)</th>
<th>RR of male breast cancer (95% CI)</th>
<th>RR of Prostate cancer</th>
<th>RR of pancreatic cancer (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>BOADICEA</td>
<td>0.89</td>
<td>1</td>
<td>1</td>
<td>age&lt;20</td>
<td>age&lt;30</td>
<td>1.82 age&lt;65 3.10 age&lt;65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (3.0146+0.02412x age)</td>
<td>20 sage&lt;29</td>
<td>30 sage&lt;39</td>
<td>4 sage&lt;40</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (6.0707−0.07775x age)</td>
<td>30 sage&lt;39</td>
<td>4 sage&lt;40</td>
<td>4 sage&lt;40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EOC model</td>
<td>0.00794</td>
<td>(4.2511−0.03226x age)</td>
<td>40 sage&lt;40</td>
<td>5 sage&lt;50</td>
<td>7 sage&lt;70</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (4.2086-0.03141x age)</td>
<td>50 sage&lt;50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA2</td>
<td>BOADICEA</td>
<td>0.96</td>
<td>1</td>
<td>1</td>
<td>age&lt;40</td>
<td>80</td>
<td>7.33 age&lt;65 5.54 age&lt;65</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (3.2153−0.008815x age)</td>
<td>20 sage&lt;29</td>
<td>4 sage&lt;40</td>
<td>4 sage&lt;40</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (4.28945−0.04462x age)</td>
<td>30 sage&lt;39</td>
<td></td>
<td>8 sage&lt;8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EOC model</td>
<td>0.0002</td>
<td>(3.96865−0.0366x age)</td>
<td>40 sage&lt;40</td>
<td>5 sage&lt;50</td>
<td>7 sage&lt;70</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (1.8169+0.00635x age)</td>
<td>50 sage&lt;50</td>
<td>4.4</td>
<td>7 sage&lt;70</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (−0.2606+0.0416x age)</td>
<td>60 sage&lt;60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13.991</td>
<td>70 sage&lt;79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PALB2</td>
<td></td>
<td>0.0064</td>
<td>0.92</td>
<td>1</td>
<td>age&lt;20</td>
<td>age&lt;30</td>
<td>1 1 age&lt;30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9.1</td>
<td>20 sage&lt;24</td>
<td>2.91 (1.40 to 6.04)</td>
<td>age&lt;30</td>
<td>7.34 (1.28 to 42.10) age&lt;30 2.37 (1.24 to 4.50) age&lt;30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8.97</td>
<td>25 sage&lt;29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8.85</td>
<td>30 sage&lt;34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8.54</td>
<td>35 sage&lt;39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8.02</td>
<td>40 sage&lt;44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7.31</td>
<td>45 sage&lt;49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6.55</td>
<td>50 sage&lt;54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.92</td>
<td>55 sage&lt;59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.45</td>
<td>60 sage&lt;64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.1</td>
<td>65 sage&lt;69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.82</td>
<td>70 sage&lt;74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.56</td>
<td>75 sage&lt;79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHEK2</td>
<td></td>
<td>0.00373</td>
<td>0.98</td>
<td>1</td>
<td>age&lt;20</td>
<td>1</td>
<td>1 1 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (1.6053−0.01483x age)</td>
<td>age&lt;20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATM</td>
<td></td>
<td>0.0018</td>
<td>0.94</td>
<td>2.10 (1.17 to 2.57)</td>
<td>1</td>
<td>1</td>
<td>1 1 1</td>
</tr>
<tr>
<td>BARD1</td>
<td></td>
<td>0.00043</td>
<td>0.89</td>
<td>2.09 (1.35 to 3.23)</td>
<td>1</td>
<td>1</td>
<td>1 1 1</td>
</tr>
<tr>
<td>RAD51C</td>
<td></td>
<td>0.00035</td>
<td>0.78</td>
<td>1.97 (1.48 to 2.62)</td>
<td>age&lt;30</td>
<td>1</td>
<td>1 1 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (−1.7974+0.07631x age)</td>
<td>30 sage&lt;60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (9.7592−0.11163x age)</td>
<td>age&lt;60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD51D</td>
<td></td>
<td>0.00035</td>
<td>0.86</td>
<td>1.82 (1.34 to 2.47)</td>
<td>age&lt;30</td>
<td>1</td>
<td>1 1 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (−2.8862+0.09965x age)</td>
<td>30 sage&lt;58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>exp (5.99144−0.05651x age)</td>
<td>age&lt;58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRIP1</td>
<td></td>
<td>0.00071</td>
<td>0.95</td>
<td>3.41 (2.12 to 5.54)</td>
<td>1</td>
<td>1</td>
<td>1 1 1</td>
</tr>
</tbody>
</table>

*Allele freq* is the pathogenic variant allele frequency in the general population. RR is relative to the general population. The BOADICEA model includes the effects of BRCA1, BRCA2, PALB2, CHEK2, ATM, BARD1, RAD51C and RAD51D, while the EOC model includes the effects of BRCA1, BRCA2, RAD51D, RAD51C, BRIP1 and PALB2. The updated parameters are the allele frequencies for PALB2, CHEK2, ATM, RAD51C, RAD51D, BARD1 and BRIP1, the SS for pathogenic variants for all genes, the RR for female breast cancer for ATM, BARD1, RAD51C and RAD51D, and the EOC, male breast cancer and the pancreatic cancer RRs for PALB2. All other parameters are as previously published. BOADICEA, Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm; EOC, epithelial tubo-ovarian cancer; RR, relative risk; SS, screening sensitivity.
Cancer genetics

risks less than 17%, (2) moderate risk, for risks in the range of 17%–30% and (3) high risk, for risks of 30% or greater. Reclassification was considered based on questionnaire-based RFs (QRFs) (RFs other than mammographic density), mammographic density (MD, based on the BI-RADS system) and a polygenic risk score (PRS). The assumed distributions and RRs for QRFs and MD have been described in detail previously.8 11 For BC, the PRS was taken to be the Breast Cancer Association Consortium 313 variant PRS, which accounts for 20% of the overall polygenic variance.9 26 For EOC, we defined three risk categories based on lifetime risk27 28: (1) near-population risk, for risks of less than 5%, (2) moderate risk, for risks in the range of 5%–10% and (3) high risk, for risks of 10% or greater, and reclassification was considered based on RFs and PRS. For EOC, the PRS was taken as the Ovarian Cancer Association Consortium 313 variant PRS, which accounts for 5% of the overall polygenic variance.11 28

Updates to tumour pathology

Both models incorporate data on BC tumour pathology, specifically ER and TN. The distribution of pathology for affected carriers of PVs differs substantially from that in non-carriers for several genes, so that pathology data can affect the carrier probabilities and hence cancer risks.11 12 In BOADICEA and the EOC model, breast tumours are classified into five groups based on ER and TN status: ER unknown, ER-positive, ER-negative/TN unknown, ER-negative/not TN and TN. Previously, the models achieved this using age-dependent distributions in the general population and BRCA1 and BRCA2 PV carriers and an age-independent distribution for CHEK2 PV carriers.12 21 Due to the lack of data, the tumour ER distribution for carriers of PV in other genes was assumed to be the same as the general population. Here, the models have been updated to incorporate age-dependent ER and TN tumour distributions for carriers of PVs in the BC susceptibility genes PALB2, CHEK2, ATM, BARD1, RAD51C and RAD51D, using data from BRIDGES.19

Continuous risk factors

The previous versions of the models included reproductive, lifestyle, hormonal and anthropometric RFs.8 11 One limitation of these models was that the RFs needed to be coded as categorical variables. Some RFs are naturally continuous, requiring prior discretisation to a finite number of categories, resulting in some loss of information and reduction in risk discrimination. Here, the methodology was extended to allow the inclusion of continuous RFs. The key challenge is to calculate the baseline incidences \( \lambda_0(t) \) in equation 1.1 (online supplemental file 1) from the population incidence and the RF distributions. The baseline incidences are calculated sequentially for each age \( t \) (considered discrete) using the values at age \( t - 1 \), starting from age 0, requiring the evolution with age of the probability distribution of those who are disease free.10  For discrete factors/gene, this involves summing over all possible categories/genotypes, but for continuous factors/genes, it would involve integrating over all possible values. In principle, these integrals could be computed (either analytically or numerically). However, at each age, the number of terms in the integrand increases by a factor of 2, so by age 80, there are \( >10^{24} \) terms, with evaluation becomes impracticable. Alternatively, the RF could be discretised into a very large number of categories. This would give a very close approximation to the continuous distribution, but (particularly once multiple RFs are considered, as here) the large number of categories would also make the calculations impractical. Instead, we propose an alternate approach in which the continuous factors are discretised with categories adapted according to the observed RF. The approach is as follows:

1. First, discretise the range of possible RF values into a finite number \( n \) of bins and calculate the probability mass and RR for each bin from the probability density and RR function for the continuous RF. This part is identical to the standard approach for discretising RFs, used in the existing models.5 For a RF, \( x \), with probability density \( P(x) \) and relative risk \( RR(x) \), the probability mass for bin \( i \) with range \([l_i, u_i]\) is:

\[
P(i) = \int_{l_i}^{u_i} P(x) \, dx
\]

and the corresponding RR is

\[
RR(i) = \frac{1}{P(i)} \int_{l_i}^{u_i} RR(x) \, P(x) \, dx
\]

2. Create an additional \( n + 1 \)th bin based on the individual’s measured RF value that has an infinitesimal width. The RR for this bin is taken as the RR at the measured value, and it has zero mass. As this bin is infinitesimal, its overlap with the other bins is zero, so there is no double-counting. This procedure creates a categorical RF with \( n + 1 \) categories, where the individual is assigned to the \( (n + 1) \)th category defined in step 2. This allows the exact value of the risk for the individual to be used, while the number of categories required to compute the baseline rates is fixed, limiting the computation time.

The accuracy of the approximation in the procedure relies on the assumption that the range of values within each bin have similar RRs, which should be reflected in the choice of discretisation scheme and the number of bins \( n \). These choices will depend on the shape of the distribution and the RR function.

The above procedure can be applied to any RF distribution or RR function. However, the process assumes that an individual’s position within the distribution is fixed with respect to age, although the value of the RF and RR may vary with age. Here, the method was applied to height.

Updates to population incidences

The baseline incidences in equation s.1 in online supplemental file 1 are birth year and country specific as a consequence of using birth year and country-specific population incidences in the constraining process. We refined the derivation of cohort-specific population incidences to account for variability in the incidences due to small numbers. In addition, we have updated existing incidences in the model to include more recent calendar periods and adapted the model to use cancer incidence from four new populations: the Netherlands, France, Slovenia and Estonia. Details are included in the online supplemental material.

RESULTS

Rare moderate-risk pathogenic variants

Table 1 summarises the models’ genetic parameter estimates, including those for the new genes. The estimated cumulative age-specific BC risks for BARD1, RAD51C and RAD51D PV carriers in BOADICEA and EOC risks for PALB2 carriers, assuming the UK incidences applicable to those born in the 1980s, are shown in figure 1. The estimated average lifetime BC risks for PV carriers are 24%, 22% and 21% for BARD1, RAD51C and RAD51D PV carriers, respectively. The estimated lifetime EOC risk for PALB2 carriers is 28%.
carriers is 5.0%. Based on the assumed allele frequencies, 0.22% of the population carry PV in the genes BARD1, RAD51C or RAD51D, and these explain on average 0.31% of the female BC polygenic variance (averaged over all ages and cohorts, weighted by the age-specific and cohort-specific BC incidences). Approximately 0.13% of the population carry PVs in PALB2, explaining 0.16% of the EOC polygenic variance and 2.5% of the male BC polygenic variance.

Figure 2A–F and online supplemental table S1 show the distributions of lifetime BC risks for carriers of PVs in BARD1, RAD51C and RAD51D for a female with unknown FH and a female whose mother is affected at age 50 based on PV carrier status alone and including QRF, MD and a PRS. Based solely on PV carrier status, all females with unknown FH would be classified as at moderate risk. When information on QRF, MD or PRS is known, there is significant reclassification to near-population and high-risk categories, which is greatest when all factors are used in combination. For example, based on lifetime BC risks and using the full multifactorial model incorporating QRF, MD and with unknown FH would be classified as at moderate risk. When information on QRF, MD or PRS is known, there is significant reclassification to near-population and high-risk categories, which is greatest when all factors are used in combination. For example, based on lifetime BC risks and using the full multifactorial model incorporating QRF, MD and PRS, 33.9% of BARD1 PV carriers with unknown FH would be reclassified from moderate risk to near-population risk, and 21.9% would be reclassified to high risk (online supplemental table S1). Similarly, BARD1 PV carriers with an affected first-degree relative would be considered high risk (risk of 33.7% by age 80) based on family history and PV status alone. Incorporating the other risk factors would reclassify 12% as near-population risk and 40.2% as moderate risk (online supplemental table S1).

Figure 2G,H and online supplemental table S2 show the distribution of lifetime EOC risks for carriers of PVs in PALB2 for a female with unknown FH and a female whose mother is affected at age 50, as a function of the RFS and PRS. For a PALB2 carrier with unknown FH, when the RFS and PRS are considered jointly, 62.4% are classified as near-population risk, 34.9% as moderate risk and 2.7% as high risk. The corresponding proportions with an affected mother are 11.2%, 55.8% and 33%, respectively. However, even among PALB2 carriers with an affected mother, 97.5% will have risks of less than 3% by age 50 (online supplemental table S2).

**Tumour pathology**

Figure 3 and online supplemental tables S3 and S4 show the age-specific distributions of ER-negative tumours and TN tumours among ER-negative tumours used in the models for PALB2, ATM, CHEK2, BARD1, RAD51C and RAD51D PV carriers based on the BRIDGES data. BARD1, RAD51C and RAD51D PV carriers predominantly develop ER-negative BCs, and the proportions decrease with increasing age. On the other hand, CHEK2 and ATM carriers primarily develop ER-positive BCs, and the proportion of ER-positive tumours increases with age. Among those with ER-negative tumours, most tumours are TN for PV carriers in all genes, except CHEK2 carriers, in whom the majority are ER-negative but not TN.

Using the updated age-specific and gene-specific ER-negative and TN tumour status distributions resulted in differences in the predicted overall and gene-specific carrier probabilities by different tumour pathology and age (figure 4). For ATM, the carrier probabilities for ER-negative tumours are reduced relative to previous estimates, reflecting the stronger association with ER-positive disease. Carrier probabilities for CHEK2 now show a decline with age for ER-negative tumours (previously, this was only predicted for ER-positive disease). The carrier probabilities for PALB2 remain similar to previous estimates. For the new genes BARD1, RAD51C and RAD51D, the carrier probabilities are, as expected, higher for ER-negative and TN diseases, but there is little variation by age.

**Continuous risk factors**

As previously, adult female height was assumed to be normally distributed with mean 162.81 cm and SD 6.452 cm, and be associated with a log-RR per SD, for both BC and EOC, of 0.10130.8.11 We therefore discretised the normal distribution such that the probability masses of the bins were given by a binomial distribution \( B \left( n - 1, \frac{1}{2}\right) \), giving sufficient discretisation to adequately capture the tails of the distribution. We examined the relative discretisation error of the predicted lifetime risk as a function of the number of bins (figure 5E,F) and chose \( n = 5 \), as the lowest number of bins such that the root-mean-square relative error was less than \( 10^{-4} \). Compared with the discrete (five-level) RF, the variance of the RR of both BC and EOC increased from 0.002 to 0.010 when height was included as a continuous RF. The effects on predicted lifetime risks are shown in figure 5A–D. Under the continuous implementation here, the lifetime...
Figure 2  Predicted lifetime cancer risks (from age 20–80 years) for a female born in 1985 with a pathogenic variant in BARD1 (breast cancer risk), RAD51C (breast cancer risk), RAD51D (breast cancer risk) and PALB2 (ovarian cancer risk) on the basis of the different predictors of risk (pathogenic variant (PV) status, questionnaire-based risk factors (QRFs), mammographic density (MD) and polygenic risk score (PRS)). All figures show the probability density against the absolute risk. Figures (A), (C), (E) and (G) show risks for a female with unknown family history, while Figures (B), (D), (F) and (H) show risks where the individual’s mother has had cancer at age 50. The backgrounds of the graphs are shaded to indicate the risk categories. For breast cancer, these are the categories defined by the National Institute for Health and Care Excellence familial breast cancer guidelines: (1) near-population risk shaded in pink (<17%), (2) moderate risk shaded in yellow (≥17% and <30%) and (3) high risk shaded in blue (≥30%). For ovarian cancer, the categories are: (1) near-population risk shaded in pink (<5%), (2) moderate risk shaded in yellow (≥5% and <10%) and (3) high risk shaded in blue (≥10%). Predictions were based on UK cancer incidences. The line-labelled population denotes the average population risk in the absence of knowledge of family history, PV status, RFs or a PRS. All figures assume the population distributions of QRFs and MD.
BC risk varied from 9.7% for the first percentile to 14.6% for the 99th percentile, whereas under the previous discrete distribution, the risks range from 10.1% to 14.2%.

DISCUSSION

This work has extended the multifactorial BOADICEA BC and EOC risk prediction models (BOADICEA v6 and the Ovarian Cancer Model v2), employing a synthetic approach. The explicit effects of PVs in RAD51C, RAD51D, BARD1 and PALB2, which have now been established as BC and/or EOC susceptibility genes and are commonly included on cancer gene panels, are now included in the models. The models have also been extended to accommodate continuous RFs, and parameterisation of tumour pathology and cancer incidence have been updated with more recent data. These represent the most comprehensive models for BC and EOC and will allow more complete BC and EOC risk assessment of those undergoing gene-panel testing. In a separate study, the BOADICEA v6 breast cancer model presented here has been validated in an independent prospective study of 66,415 women attending mammographic screening in Sweden. The full model, including RFs, mammographic density, PRS and PVs in BRCA1, BRCA2, PALB2, CHEK2, ATM, BARD1, RAD51C and RAD51D was well calibrated overall (calibration slope 0.97 (95% CI: 0.95 to 0.99)) and in deciles of predicted 5-year risks and had a C-index of 0.71 (95% CI: 0.68 to 0.74) for discriminating between affected and unaffected women.

By explicitly modelling the effects of PVs in the new cancer susceptibility genes, the models provide personalised cancer risks of PV carriers when combined with QRFS, MD and PRS. Although the numbers affected by these changes will be small at population level, for individuals with RAD51C, RAD51D and BARD1 PVs and their families, the updated risks will be clinically important. RAD51C, RAD51D and BARD1 (like ATM and CHEK2) would be classified as ‘moderate risk’ BC genes based on the average risks. However, according to the BOADICEA predictions, over half (56%–59%) of carriers of PVs in these genes in the population would be reclassified from being in the moderate BC risk category to either being near-population risk (34%–44%) or high risk (15%–22%), if data on the other RFs were incorporated (online supplemental table S1). Such changes may have important implications for discussions around earlier or more frequent screening or on risk-reduction options for these women. Similarly, based on the multifactorial EOC model —38% of PALB2 PV carriers will have lifetime EOC risks of >5% (online supplemental table S2), which may influence recommendations on the timing of risk-reducing surgery.

As previously, the models assume that the effects of the PVs in the new genes interact multiplicatively with the PRS and the RFs. No studies have yet assessed the joint effects of PVs in these genes and the PRS or RFs. Previous results for CHEK2 and ATM suggest that the multiplicative model holds true for earlier versions of the PRS. Unlike CHEK2 and ATM, however, the new genes predispose more strongly to ER-negative disease, and the combined effect may depart from the multiplicative assumption. Demonstrating this explicitly for the new genes will be challenging given the rarity of the mutations. The multiplicative model has also been shown to be reasonable for the combined effects of PRS and RFs, but there is as yet no large-scale evaluation of the combined effects of PVs and RFs. However, recent prospective validation studies of the current and previous versions of the models suggest that, overall, the models fit well. Should deviations from the multiplicative model between these PVs and RFs emerge, the model can be updated to take them into account.

Both the BC and EOC models incorporate PVs’ effects using the estimated population allele frequencies and RRAs. These are combined with reference population incidences to calculate absolute risks while constraining the overall incidences over the RFs included in the model. Our implementation used RR and allele frequency estimates from the largest available studies on those of European ancestry. These were assumed to be constant across all countries. Available data are currently too sparse to obtain country-specific estimates. Although there is no evidence that RRAs vary among populations, the allele frequencies are likely to vary to some extent. This is most apparent for CHEK2, where the founder c.1100delC variant (p.Thr367Metfs*15) is common in northwest Europe with carrier frequencies between 0.3% and 1.2% and explains the majority of carriers but is rare or absent in other populations. If population-specific variant frequencies can be generated, the model can be easily updated to accommodate these. Nevertheless, by allowing population incidences to vary by country, the predicted absolute risks given by the models are country-specific.

The updated age-specific distributions of tumour ER and TN status for six of the BC susceptibility genes in the model...
Cancer genetics

Figure 4  The probabilities of carrying a pathogenic variant estimated by BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) model in the genes PALB2, CHEK2, ATM, BARD1, RAD51C and RAD51D for an affected female born in 1985 as a function of her age at diagnosis based on different tumour pathology. Figures (A), (C), (E) and (G) show the probabilities based on the updated proportions (current model), while figures (B), (D), (F) and (H) are based on the previously assumed tumour pathology proportions (previous model version) and where proportions for BARD1, RAD51C and RAD51D, which were not in the previous model, are assumed to be the same as in the general population. In figures (A) and (B), the woman has had an oestrogen receptor-positive (ER+) tumour; in figures (C) and (D), the female has had an oestrogen receptor-negative (ER−) tumour, but the triple-negative (TN) status is unknown; in figures (E) and (F), the woman has had an ER− tumour that is not TN and in figures (G) and (H), the woman has had a TN tumour. Predictions are based on UK cancer incidences. BC, breast cancer.
Cancer genetics

(PALB2, CHEK2, ATM, BARD1, RAD51C and RAD51D) should allow better differentiation between PVs that may be present in a family and provide age-specific and gene-specific mutation carrier probabilities consistent with the prevalence of PVs observed in Mavaddat et al. We note, however, that estimates are more uncertain at very young and very old ages, where the data are sparse, and more extensive validation may be required in these age-groups. Since PV carrier probabilities are used internally in the models, these will also impact the predicted absolute risks for all unaffected individuals if information on tumour characteristics is available for affected relatives whether or not they carry a PV.

We have developed a novel methodological approach for including continuous RFs into the models. We demonstrated this by including height in both the BC and EOC models, allowing for more nuanced predictions and improving the risk discrimination. While the resulting discrimination based on height alone is modest, the framework will allow other more predictive RFs to be included in the model if accurate risk estimates become available. The most important example is MD: continuous measures of MD, available through tools such as STRATUS, CUMULUS and Volpara, have been shown to have stronger associations with BC risk than the categorical BI-RADS system. Other examples include BMI and ages at menarche and menopause. Further, the method could be applied to the joint distribution of several continuous risk factors, where the integrals in equations (1) and (2) become multidimensional integrals.

We have further refined the method for creating cohort incidences from calendar period incidences (online supplemental material). The approach provides incidences that are less sensitive to year-on-year fluctuations by averaging over all years in the birth cohort. This method is particularly useful for cancers with low incidences, such as EOC and male BC, where the population size is small, and there is no prior averaging over calendar

Figure 5 Predicted lifetime breast and ovarian cancer risks as a function of height for a female born in 1985 with unknown family history, comparing the updated model, where height is treated as continuous, to the previous model, where height was treated as categorical. Figures (A), (C) and (E) show breast cancer, while figures (B), (D) and (F) show ovarian cancer risks. Figures (A) and (B) show the predicted risk as a function of height, while figures (C) and (D) show the probability density/mass of risk as a function of height. Predictions are based on UK cancer incidences. Figures (E) and (F) show the log (base 10) of the root-mean-squared relative discretisation error as a function of the number of bins. The error was taken to be the absolute difference between the value and the asymptotic extrapolation of the measurements as a function of the number of bins. The average is taken over 100 heights that are spaced 1% apart, from 0.5% to 99.5%.

years. The refinement will have little effect on incidences from larger countries.

Our models have certain limitations. No single dataset containing all the required information was available to construct the multifactorial models, so the models were extended via a synthetic approach. The new model parameters were taken from extensive, well-designed published studies together with existing parameters from model fitting. 8 We and others have used this approach for developing previous versions of the models, 9 11 12 21 40 41 which have been shown to provide clinically valid predictions. 42 43 As is the case for the previous versions, the updates presented here are primarily based on studies of those of European ancestry in developed countries. There is little evidence that the RRs associated with PVs differ by ancestry. The PV frequencies are also broadly similar across populations, except for specific founder mutations and CHEK2 PVs, which have a much higher frequency in European than non-European populations. However, other parameters in the model, including RF and PRS distributions, will differ by population, and the model will need to be adapted for use in non-European ancestry populations and developing countries. The synthetic approach presented here allows the model to be easily customised to other populations as better estimates become available. 44 45 Although we used the associations between PVs and tumour ER and TN status, the models do not currently consider the associations with intrinsic BC subtypes based on combinations of ER, progesterone receptor, HER2 and/or grade. 19 The methodology described here could be used to further extend the models to consider these BC subtypes. Finally, the models make the simplifying assumption that PVs in the assumed BC and EOC susceptibility genes are associated with similar risks to those for truncating variants. These would include missense variants which have similar risks to truncating variants. However, there is evidence that missense variants in CHEK2 and ATM are associated with BC risk, which may be different from the risks for truncating variants. 44 The models would not be applicable to carriers of such variants.

The new model features have been built on the established and well-validated BOADICEA and EOC models. 8 11 42 The updated models will allow for more personalised risk assessment and can help guide decisions on screening, prevention, risk factor modification or other risk-reducing options. The models presented are now available for use by healthcare professionals through the user-friendly CanRisk webtool (www.canrisk.org, CanRisk V2).

Author affiliations
1Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
2Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
3Wolfson Institute of Population Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK
4Department of Medical Genetics and National Institute for Health Research, Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK
5Centre Hospitalier Universitaire de Québec–Université Laval Research Center, Université Laval, Quebec, Quebec, Canada
6Division of Molecular Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands
7Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
8Epidemiology and Cancer Registry, Institute of Oncology, Ljubljana, Slovenia
9Clinic of Hematology and Oncology, Tartu University Hospital, Tartu, Estonia
10Service de Génétique, Institut Curie, Paris, France
11Department of Epidemiology, Netherlands Cancer Institute, Amsterdam, The Netherlands
12Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK

Contributors Conceptualisation: ACA, PPDP, DFE; Methodology: AL, ACA, DFE, PPDP; Software: AL, AC, TC, LE, SA, FMW, MT, JR, JJU-S, ACA; Investigation: AL, ACA, PPDP, DFE; Formal analysis: AL, NM, LF; Funding acquisition: ACA, MKS, PD, JS; Resources: PPDP, DFE, ACA, VZ, HJ, EM-F, ADP, MR, TMM; Supervision: ACA; Visualisation: AL; Writing—original draft: AL, DFE, ACA; Writing—review and editing: all authors.

Funding This work has been supported by grants from Cancer Research UK (C12292/A20861 and PRPPIGM-Nov20100002); the European Union’s Horizon 2020 research and innovation programme under grant agreement numbers 633784 (B-CAST) and 634935 (BRIDGES); the PERSPECTIVE 18i project, which the Government of Canada funds through Genome Canada (#13529) and the Canadian Institutes of Health Research (#155865), the Ministère de l’Économie et de l’Innovation du Québec through Genome Québec, the Quebec Breast Cancer Foundation, the CHU de Quebec Foundation and the Ontario Research Fund. MT is supported by the NIH/NCI Biomedical Research Centre (BCR-1215-20014).

Competing interests The authors ACA, DFE, AL, AC and TC are named inventors of BOADICEA v5 commercialised by Cambridge Enterprise. AL is now employed by Illumina.

Patient consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study. The models are implemented in the CanRisk tool, which is freely available at www.canrisk.org.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

ORCID iDs Andrew Lee http://orcid.org/0000-0003-0677-0252
Peter Devilee http://orcid.org/0000-0002-8023-2009
Paul PD Pharoah http://orcid.org/0000-0001-8494-732X
Antonis C Antoniou http://orcid.org/0000-0001-9223-3116

REFERENCES
Cancer genetics


Supplementary Material

Pathogenic Variants in Cancer Susceptibility Genes

Model definition

The BOADICEA and epithelial tubo-ovarian cancer (EOC) models assume that the cancer incidences for individual i at age t, $\lambda^{(i)}(t)$, depend on their underlying genotype through a model of the form:

$$
\lambda^{(i)}(t) = \lambda_0(t) \exp \left( \sum_{\mu=1}^{N_{MG}+1} \left[ \beta_{MG\mu}(t) + \sum_{\rho} \beta_{RF\rho}(t) \cdot z_{RF\rho}^{(i)} \right] \prod_{v=1}^{\mu-1} \left[ 1 - G_{v}^{(i)} \cdot g_{v}^{(i)} \right] \right) + \beta_{PG}(t) x_{PG}^{(i)},
$$

(s. 1)

where $\lambda_0(t)$ is the baseline incidence (applicable to a non-PV carrier with a zero polygenotype and unknown RFs). $N_{MG}$ is the number of major genes present in the model, which for the previous versions of both models was five. $G_{v}^{(i)}$ are indicator variables for the presence/absence of a PV in a major gene in person i, taking values 1 if a PV is present and 0 otherwise with $\mu = 1, ..., N_{MG}$ representing the genes present in the model in the dominance order and $\mu = N_{MG} + 1$ corresponding to non-carriers of PVs, where $G_{N_{MG}+1}^{(i)} = 1$ for non-carriers of any PV and 0 otherwise. The cancer incidences associated with homozygous and heterozygous carriers of PVs in each gene are assumed to be the same, and the risk to carriers of PVs in more than one gene is assumed to be that of the higher-ranked PV in the dominance order. Because PVs are rare, this model can be well approximated by assuming a single locus of PVs in more than one gene is assumed to be that of the higher-ranked PV in the dominance order. Because PVs are rare, this model can be well approximated by assuming a single locus with $N_{MG} + 1$ alleles, one representing the presence of a PV in each of the $N_{MG}$ genes and an additional wild-type allele representing absence of PVs in all genes. $\beta_{MG\mu}(t)$ represent the age-specific log-relative risks (log-RRs) associated with the major genes relative to the baseline incidence. The relative risks (RR) assumed for the major genes are summarised in Table 1. $x_{PG}^{(i)}$ is the polygenotype for individual i, assumed normally distributed in the general population with mean 0 and standard deviation 1, and $\beta_{PG}(t)$ is the age-specific log-RR per standard deviation associated with the polygene, relative to the baseline incidence. When a PRS is known, the polygenotype is decomposed into an observed and residual component where the observed component is given by the PRS, $\rho$ indexes the RFs that are present in the model, which are modelled as categorical factors. $\beta_{RF\rho}(t)$ is the vector (length $\kappa_{\rho} - 1$ were $\kappa_{\rho}$ is the number of categories for RF $\rho$, with one category being the baseline) of age-specific log-RRs associated with RF $\rho$, which may depend on the major genotype $\mu$, and $z_{RF\rho}^{(i)}$ is the corresponding vector of indicator variables (0 or 1) that indicate the category of RF $\rho$ for individual i (1 for the observed category, 0 otherwise, with all elements 0 for the baseline). The baseline incidences $\lambda_0(t)$ are determined so that the total age-specific incidences, summed over the RFs and genotypes, agree with the population incidence (given the assumed population distributions and RRs). The population incidences are birth-cohort and country-specific, but this dependence is omitted from equation (s.1) for clarity of notation. The RRs and distributions of the RF have been described elsewhere. To allow appropriately for
missing RF information, only those RFs measured on a given individual are considered (thus, the baseline incidence, $\lambda_0(t)$, are determined for each individual dependent on their measured RFs).

The models assume that RRs associated with PVs in the major genes are log-additive (multiplicative) with the RFs and the polygenic component. The model also assumes that the PVs and the PRS combine multiplicatively (conditional on other factors).

The models evaluate pedigree likelihoods using the MENDEL software. As MENDEL considers only finite discrete genotypes, the polygenotype is approximated by the hypergeometric polygenic model.

Both models consider family history of breast cancer (BC), EOC, pancreatic cancer (PaC) and prostate cancer (PrC). The incidences of each cancer are assumed independent, conditional on the genotypes and RFs in the model. In BOADICEA, EOC, PaC and PrC are assumed to depend only on the major genotype. Correspondingly, in the EOC model, BC, PaC and PrC are assumed to depend only on the major genotype.

Adjusting the residual polygenic component after the inclusion of new major genes

The variance due to PVs in each gene at age $t$ is given by:

$$\text{var}(t, \mu) = \log \left( \frac{\left(1 - f_{\mu}\right)^2 + f_{\mu}(2 - f_{\mu}) \exp(2\beta_{MG\mu}(t))}{\left(1 - f_{\mu}\right)^2 + f_{\mu}(2 - f_{\mu}) \exp(\beta_{MG\mu}(t))} \right),$$

where $f_{\mu}$ is the population allele frequency of gene $\mu$; the variance components are assumed to be additive. This process also considered the updated RR and PV frequencies for the previously included genes. For BOADICEA, the overall BC polygenic variance was $4.83 - 0.5961 \times t$ for females and $1.4$ for males, while for the EOC model, the overall EOC polygenic variance was $1.434$.

Allele Frequencies

Allele frequencies for all genes, except BRCA1 and BRCA2, were taken from the BRIDGES study. The frequencies were based on the frequency of protein-truncating variants in European ancestry controls. To account for the incomplete sensitivity of the sequencing as performed in BRIDGES, the frequencies were adjusted by dividing by $c s(1 - v)$, where $c$ is the proportion of the coding sequence of each gene determined to be callable, $s$ is the proportion of variants in the called sequence across all genes that were detected (estimated to be 0.957), and $v$ is the proportion of the pathogenic variants expected to be copy variants. For CHEK2, the adjustment was applied to variants excluding c.1100delC. Details are given in the Supplementary Material of Dorling et al. For BRIP1, $v$ was assumed to be 0.05. The BRCA1 and BRCA2 frequencies from the previous versions of BOADICEA and the EOC model were used for consistency.

Sensitivities

The default sensitivities are based on the assumption that protein truncating variants and known pathogenic missense variants are detected with close to 100% sensitivity in clinical tests but that, except for BRCA1 and BRCA2, large rearrangements are not detected. The
sensitivities are therefore given by \( 1 - \nu \), as above. For \( BRCA1 \) and \( BRCA2 \), sensitivities were defined by assuming that the main source of insensitivity was missense variants not classified as pathogenic – the frequencies of these variants have been estimated by Dorling et al.\(^{10}\). Were large re-arrangements not tested for, the corresponding sensitivities for \( BRCA1 \) and \( BRCA2 \) would be reduced to \( \sim 76\% \) and \( 95\% \) respectively (owing to the much higher frequency of large re-arrangements in \( BRCA1 \)).

**BRCA2**: ovarian cancer relative risks updates

Previous estimates of the EOC relative risks for \( BRCA2 \) PV carriers were obtained during the BOADICEA model fitting process, using complex segregation analysis in families with \( BRCA2 \) PVs.\(^{2}\). This involved fitting models in which the log-relative risks were piecewise linear functions of age. Due to the very small number of EOCs diagnosed in ages 65 years and over in the original dataset, the RR was estimated to decrease rapidly from 23.7 at age 58, to 1.59 at age 69 and remain constant at that level thereafter. However, more recent data suggest that the EOC RR for ages 70 and over are higher.\(^{11}\). The original RR estimate of 1.59 may result in an underestimation of risks for older \( BRCA2 \) carriers. We therefore updated the log-RR function included in the model by re-deriving the piecewise log-RR linear function such that the EOC RR decreases less rapidly from 23.7 at age 58 to 4.4 for ages 70 and over. The RR=4.4 estimate used for ages 70 and over was obtained from a prospective cohort analysis of \( BRCA2 \) PV carriers.\(^{11}\).

The updated log-RR EOC parameters for ages 58 and over for \( BRCA2 \) carriers are shown in Table 1 and the resulting age-specific EOC cumulative risks are shown in Figures 3.

**Population Incidences**

The BOADICEA and EOC models both allow population customisation via population-specific incidences.\(^{4,6,12}\). Here the models are extended with incidences from the Netherlands, France, Slovenia and Estonia. Incidences for the Netherlands were taken from Statistics Netherlands for 1950-1988 and the Netherlands Cancer Registry for 1989-2017, where BC incidences exclude ductal carcinomas in situ, as these are not included in the models.\(^{13,14}\). Incidences for France were taken from CI5Plus and CI5 for 1977-1989 using nine registries and from INCa/Santé Public France for 1990-2018.\(^{15,17}\). Incidences for Slovenia covering 1961-2016 were taken from the Slovenian Cancer Registry.\(^{18}\). Incidences for Estonia covering 1968-2018 were taken from the Estonian National Institute for Health Development.\(^{19}\). Predicted lifetime breast and EOC risks using these incidences are shown in Figure s1.

Incidence for some of the existing regions were updated using data from more recent calendar years. For the UK, incidences covering 2011-2017 were added.\(^{20}\). For Denmark, Finland, Iceland, Norway and Sweden, incidences covering 2011-2018 were added.\(^{21,22}\). For Australia, incidences covering 2011-2017 were added.\(^{23}\). For the USA, incidences covering 2013-2018 from 21 registries were added.\(^{24}\). For New Zealand, incidences covering 2010-2018 were added.\(^{25,26}\). For Canada, incidences covering 2011-2018 were added.\(^{27}\). Figure s2 (a) shows the updated incidences’ effects on the cohort incidences for UK female breast cancer incidences for those born in the 1980s.
The models use calendar-specific population incidences to calculate cohort-specific incidences, where the cohorts are defined by decadal birth year ranges (1910-1919, 1920-1929, 1930-1939, 1940-1949, 1950-1959, 1960-1969, 1970-1979 and 1980-1989 with individuals born before/after the first/last cohort, assumed to have the same incidences as the first/last cohort). The original model used UK incidences from CI5, which reported calendar incidences averaged in 5-year calendar-period bins. Cohort incidences were then taken as those for someone born in the middle year of each range to represent that cohort (1915 for 1910-1919 etc.). However, some of the other regions have smaller populations and report annual-calendar-period specific incidences. For these populations, especially for cancers with low incidences (e.g., EOC and male BC), using a single year to represent the cohort can lead to cohort incidences dominated by year-on-year calendar fluctuations. The methodology was refined by deriving new sets of cohort incidences. In these, the age-specific incidences for an individual in the cohort were taken as the average of the age-specific incidences applicable to those born in each year of the birth-cohort range. The average age- and cohort-specific incidences were then smoothed using LOWESS with linear regression and a bandwidth of 0.2. Figure 2 (b) shows the effects of the new averaging method on cohort incidences for Estonian male breast cancer incidences for those born in the 1920s.

Further, previously, incidences for years before/after the earliest/latest calendar year were taken to be the same as those in the earliest/latest calendar year available. Again, for regions with small populations presenting annual calendar-period incidences and cancers with low incidences, the cohort incidences can be adversely affected by statistical anomalies present in incidences of the earliest/latest calendar year. The methodology was refined with incidence for years before/after the earliest/latest calendar year taken as the average of the first/last five years of the available annual calendar-period incidences.

Algorithm optimisation

The BOADICEA future risk calculations rely on calculating pedigree likelihoods under the assumed genetic models of inheritance. The inclusion of additional genes (RAD51C, RAD51D, BARD1) in the model resulted in a substantial increase in runtime. This is further compounded by the fact that separate pedigree likelihood calculations are required for risk predictions at multiple future time-points when using the CanRisk tool (e.g. in annual, or 5-year intervals). To reduce the programme runtime when using CanRisk we re-formulated the underlying algorithm to calculate the future risks as follows.

BOADICEA calculates the probability that an individual develops breast (or ovarian) cancer over a given time period, given the age of the proband, the genotypes, other risk factors, and family history:

\[ P(D(t_1)|D(t_0), D_R, z, \theta) \]

Where \( D(t) \) is the phenotype of the proband at time \( t \), \( D_R \) represents the phenotypes of the all the relatives, \( z \) are the risk factors measured on the proband and \( \theta \) are the genetic model parameters (allele frequencies, relative risks etc). \( t_0 \) is the current age of the proband and \( t_1 \)
the future age at which the predictions are being made. In practice, these are calculated as the ratio of two pedigree likelihoods:

\[
\frac{P(D(t_1), D_R, Z, \theta)}{P(D(t_0), D_R, Z, \theta)}
\]

The numerator and denominator probabilities are the probabilities of the full set of phenotypes in the pedigree at times \(t_1\) and \(t_0\) and are calculated in MENDEL, using a pedigree peeling algorithm. When predicting future risks, this involves performing this calculation repeatedly at several time-points. However, under the standard assumption in pedigree likelihood calculations, the phenotype of the proband is conditionally independent of those of relatives given the genotypes of the relatives. Thus:

\[
\sum_G P(D(t_1)|G, D_R, Z, \theta) P(G, D(t_0), D_R, Z, \theta)
\]

which can be re-written as:

\[
\sum_G P(D(t_1)|G, D_R, Z, \theta) P(G|D_R, D(t_0), Z, \theta)
\]

where \(G\) is the full set of genotypes (including the full measured and unmeasured polygenic or major gene components) and

\[
P(G|D_R, D(t_0), Z, \theta) = \frac{P(G, D(t_0), D_R, Z, \theta)}{\sum_{G'} P(G', D(t_0), D_R, Z, \theta)}
\]

Therefore, the risk prediction (expression (s.2)) can be performed by first calculating the genotype probabilities for the proband given the phenotypes at time \(t_0\) (i.e. a single, time-consuming pedigree likelihood calculation) and then calculating the penetrance function for the proband at multiple time-points \(P(D(t_1)|G, Z, \theta)\), which does not involve any pedigree likelihood calculations.

The risk calculations under the revised and original formulations are identical, but when calculating the remaining lifetime cancer risks used in the CanRisk tool (www.canrisk.org), there is a 50-90% reduction in computation time under this revised formulation, depending on the proband’s age (Figure s4).
Figure s1. Predicted lifetime (age 20 to 80 years) breast and ovarian cancer risk by age for a female born in 1985 with unknown family history (ie average female in the population) comparing risks using incidences for the UK, France, the Netherlands, Slovenia, and Estonia. Figure (a) shows breast cancer risks, where risks for the UK, France, the Netherlands, Slovenia, and Estonia are 12.0%, 12.1%, 12.1%, 8.8% and 7.4%, respectively. Figure (b) shows ovarian cancer risks, where risks for the UK, France, the Netherlands, Slovenia, and Estonia are 1.8%, 1.2%, 1.3%, 1.1% and 1.5%, respectively.

Figure s2. Smoothed Cohort-specific population incidences. Figure (a) shows female breast cancer incidences for the UK for those born in the 1980s for the previous incidences (using incidences up to and including 2010) and for the updated incidences (using incidences up to and including 2017), where both datasets use the average over the birth years in the cohort. Figure (b) shows male breast cancer incidences for Estonia for those born in the 1920s, using incidence from a single birth year to represent the cohort (labelled 1925) and using the average over the birth years in the cohort (labelled 1920-1929).
Figure s3. Revised epithelial ovarian cancer risks for a BRCA2 pathogenic variant carrier with unknown family history using the updated BRCA2 relative risks. Figure (a) shows the cumulative risk by age, while figure (b) shows the distribution of absolute by age 80 on the basis of the different predictors of risk (pathogenic variant status (PV), questionnaire-based risk factors (QRFs), mammographic density (MD), and PRS).
Figure S4. Ratio of run-times for calculating the remaining life-time risks in the CanRisk tool under the optimised algorithm compared to the original implementation.
## New Susceptibility Genes

### Breast Cancer: BARD1, RAD51C and RAD51D

Table S1. Predicted 10-year (age 40 to 50 years) and lifetime (age 20 to 80 years) breast cancer risk for a female born in 1985 with unknown family history and for a female with a mother affected at age 50. The columns labelled “Risk” contain risks in the absence of information about questionnaire-based risk factors (QRF), mammographic density (MD) or a polygenic risk score (PRS). The columns labelled “Risk” also contain risks in the absence of information about questionnaire-based risk factors (QRF), mammographic density (MD) or a polygenic risk score (PRS). The other columns show the distribution of females based on these risk factors in the risk categories defined in the NICE familial breast cancer guidelines: 1) near-population risk, shaded pink (< 17% lifetime risk; < 3% 10-year risk), 2) moderate risk, shaded yellow (≥ 17% and < 30% lifetime risk; ≥ 3% and < 8% 10-year risk) and 3) high risk, shaded blue (≥ 30% lifetime risk; ≥ 8% 10-year risk). Column headings are shaded the same colours as the corresponding lines in Figure 2. Predictions are based on UK cancer incidences, assuming the population distributions of QRFs and MD.

<table>
<thead>
<tr>
<th>Family History</th>
<th>Risk Horizon</th>
<th>PV Screening</th>
<th>QRF</th>
<th>QRF &amp; MD</th>
<th>PRS</th>
<th>QRF &amp; PRS</th>
<th>QRF, MD &amp; PRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unaffected</td>
<td>10-Year</td>
<td>Untested</td>
<td>1.7</td>
<td>96.4</td>
<td>0</td>
<td>90.0</td>
<td>89.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BARD1</td>
<td>3.6</td>
<td>48.8</td>
<td>0.6</td>
<td>46.5</td>
<td>57.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAD51C</td>
<td>3.4</td>
<td>57.3</td>
<td>0.4</td>
<td>50.8</td>
<td>60.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAD51D</td>
<td>3.1</td>
<td>67.3</td>
<td>0.2</td>
<td>56.5</td>
<td>64.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No PV</td>
<td>1.6</td>
<td>97.1</td>
<td>0.2</td>
<td>91.5</td>
<td>91.9</td>
</tr>
<tr>
<td>Unknown</td>
<td>Lifetime</td>
<td>Untested</td>
<td>12</td>
<td>92.1</td>
<td>0.2</td>
<td>87.5</td>
<td>85.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BARD1</td>
<td>23.5</td>
<td>12.5</td>
<td>10.4</td>
<td>19.8</td>
<td>30.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAD51C</td>
<td>22.3</td>
<td>18.9</td>
<td>8.4</td>
<td>24.8</td>
<td>35.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAD51D</td>
<td>20.8</td>
<td>29.7</td>
<td>6.4</td>
<td>32.4</td>
<td>42.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No PV</td>
<td>11.6</td>
<td>92.9</td>
<td>0.1</td>
<td>89.1</td>
<td>86.7</td>
</tr>
<tr>
<td>Affected</td>
<td>10-Year</td>
<td>Untested</td>
<td>3.5</td>
<td>51.8</td>
<td>0.5</td>
<td>53.7</td>
<td>63.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BARD1</td>
<td>6.4</td>
<td>11.8</td>
<td>11.7</td>
<td>12.5</td>
<td>23.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAD51C</td>
<td>6.1</td>
<td>2.8</td>
<td>9.5</td>
<td>12.2</td>
<td>26.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAD51D</td>
<td>5.6</td>
<td>3.8</td>
<td>7.3</td>
<td>15.8</td>
<td>30.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No PV</td>
<td>3.1</td>
<td>68.7</td>
<td>0.3</td>
<td>62.2</td>
<td>72</td>
</tr>
<tr>
<td>At age 50</td>
<td>Lifetime</td>
<td>Untested</td>
<td>19.3</td>
<td>40.6</td>
<td>4.8</td>
<td>43.5</td>
<td>54.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BARD1</td>
<td>33.7</td>
<td>0.1</td>
<td>62</td>
<td>3.4</td>
<td>7.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAD51C</td>
<td>32.4</td>
<td>2.4</td>
<td>52.4</td>
<td>4.1</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAD51D</td>
<td>30.4</td>
<td>0.5</td>
<td>40.2</td>
<td>5.5</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No PV</td>
<td>17.9</td>
<td>55.3</td>
<td>3.8</td>
<td>53.5</td>
<td>62.6</td>
</tr>
</tbody>
</table>
**Ovarian cancer: PALB2**

Table S2. Predicted ovarian cancer risk to age 50 (age 20 to 50 years) and lifetime risk (age 20 to 80 years) for a female born in 1985 with unknown family history and for a female with a mother affected at age 50. The columns labelled "Risk" contain risks in the absence of information about risk factors (RF) or a polygenic risk score (PRS). The other columns show the distribution of females based on these risk factors falling into risk categories defined as: 1) near-population risk, shaded pink (< 5% lifetime risk; < 3% risk to age 50), 2) moderate risk, shaded yellow (≥ 5% and < 10% lifetime risk; ≥ 3% and < 5% risk to age 50) and 3) high risk, shaded blue (≥ 10% lifetime risk; ≥ 5% risk to age 50). Column headings are shaded the same colours as the corresponding lines in Figure 1. Predictions are based on UK cancer incidences, assuming the population distributions of QRFs and MD.

<table>
<thead>
<tr>
<th>Family History</th>
<th>Risk Horizon</th>
<th>PV Screening</th>
<th>RF</th>
<th>PRS</th>
<th>RF &amp; PRS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pop</td>
<td>Mod</td>
<td>High</td>
</tr>
<tr>
<td>Unknown</td>
<td>Risk to age 50</td>
<td>Untested</td>
<td>0.3</td>
<td>99.9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PALB2</td>
<td>0.8</td>
<td>99.9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No PV</td>
<td>0.3</td>
<td>99.9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Lifetime</td>
<td>Untested</td>
<td>1.8</td>
<td>99.9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PALB2</td>
<td>5</td>
<td>61.8</td>
<td>37.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No PV</td>
<td>1.7</td>
<td>99.9</td>
<td>0</td>
</tr>
<tr>
<td>Mother affected at age 50</td>
<td>Risk to age 50</td>
<td>Untested</td>
<td>1</td>
<td>99.9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PALB2</td>
<td>1.5</td>
<td>98.4</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No PV</td>
<td>0.5</td>
<td>99.9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Lifetime</td>
<td>Untested</td>
<td>5</td>
<td>60</td>
<td>39.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PALB2</td>
<td>9.6</td>
<td>6</td>
<td>58.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No PV</td>
<td>3.3</td>
<td>91.3</td>
<td>8.6</td>
</tr>
</tbody>
</table>
### Tumour Pathology Subtypes

<table>
<thead>
<tr>
<th>AGE</th>
<th>GENERAL POPULATION</th>
<th>BRCA1</th>
<th>BRCA2</th>
<th>PALB2</th>
<th>CHEK2</th>
<th>ATM</th>
<th>BARD1</th>
<th>RAD51C</th>
<th>RAD51D</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.4615 0.9483 0.8638</td>
<td>0.201 0.5316 0.2884 0.7706 0.7471 0.6255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S3: Age-specific proportion of oestrogen receptor-negative tumours among all female breast cancer tumours in the general population and carriers of pathogenic variants in the breast cancer susceptibility genes used in the BOADICEA model.
Table S4: Age-specific proportion of triple-negative tumours among female oestrogen receptor-negative breast cancer tumours in the general population and carriers of pathogenic variants in the breast cancer susceptibility genes used in the BOADICEA model.
References


